DOCUMENT RESUME

ED 401 317 TM 025 888

AUTHOR Breland, Hunter M.
TITLE Writing Skill Assessment: Problems and Prospects. Policy Issue Perspective Series.
INSTITUTION Educational Testing Service, Princeton, NJ. Policy Information Center.
PUB DATE Apr 96
NOTE 29p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS College Entrance Examinations; Computer Uses in Education; Cost Effectiveness; *Educational Technology; *Essay Tests; High Schools; *Performance Based Assessment; Prediction; *Test Construction; Test Reliability; Test Use; Test Validity; Writing Evaluation; *Writing Skills; Writing Tests
IDENTIFIERS *Authentic Assessment; *Free Response Test Items

ABSTRACT

Recent trends in writing skill assessment suggest a movement toward the use of free-response writing tasks and away from the traditional multiple-choice test. A number of national examinations, including major college admissions tests, have included free-response components. Most of the arguments in support of this trend relate to the hypothesized effects of testing on curriculum and instruction, but others center around systemic validity and authenticity. There are questions in these areas, however, beginning with the question of what the content of a writing assessment should be. The reliability of free-response writing tests is often reported in terms of interrater reliability, but correlations of scores assigned by different raters can inflate the estimate of reliability. Combining assessment types, essay and multiple choice, is a way to improve reliability that is proving workable. The predictive effectiveness of writing skill assessments is related to reliability. Issues of fairness, comparability, cognitive complexity, and cost and efficiency must be addressed in the construction of free-response writing skill assessments. Technology seems to be an important key to the future of writing skill assessment. The future seems to one of increasing acceptance of performance tasks, and these will be best administered through the computer. (Contains 1 figure and 51 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 401 317

A POLICY ISSUE PERSPECTIVE

POLICY INFORMATION CENTER
Educational
Testing Service

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY
R. COLEY

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

Writing Skill Assessment

Problems and Prospects

by Hunter M. Breland

BEST COPY AVAILABLE

TM025888

Writing Skill Assessment: Problems and Prospects

By

Hunter M. Breland
Educational Testing Service

April 1996

ETS
Policy Information Center
Princeton, NJ 08541

Copyright © 1996 by Educational Testing Service. All rights reserved.

Table of Contents

	Page
Preface	i
Introduction	1
Arguments for Free-Response Writing Tasks	4
Cautions from the Measurement Community	7
Conclusion	16
References	19

Preface

Recent trends in the assessment of writing signal a turn away from traditional multiple-choice tests and toward the assessment of actual writing performance. Hunter Breland draws on the experience of a wide range of programs in which writing is assessed to provide a comprehensive review of assessment practices, new and old. He describes the arguments for more authentic writing assessment as well as the important issues of validity, reliability, comparability, and fairness that must be considered. His long experience in research and development in the writing assessment areas uniquely qualifies him to extract from this experience what is useful to a nontechnical audience.

This report is in the Policy Issue Perspective series published by the Center. In this series, research and experience are combined to present both knowledge and professional judgment.

The manuscript was reviewed by Brent Bridgeman and Claudia Gentile at ETS. Gwen Shrift edited it and Carla Cooper provided desktop publishing services.

Paul E. Barton
Director
Policy Information Center

Introduction

Recent trends in writing skill assessment suggest a distinctive movement toward the use of free-response writing tasks and away from traditional multiple-choice assessments.

Recent trends in writing skill assessment suggest a distinctive movement toward the use of free-response writing tasks and away from traditional multiple-choice assessments. These trends relate to a more general movement toward performance-based assessment. Testing programs that have added free-response essay assessments recently include a number of statewide assessments, the National Assessment of Educational Progress (NAEP), the Medical College Admission Test (MCAT), and the Graduate Management Admission Test (GMAT). The Graduate Record Examination (GRE) program is planning to add at least one free-response essay. The Law School Admission Test (LSAT) introduced a free-response writing task early in 1982. Despite these trends in some programs, however, it is also clear that other testing programs are not joining this trend or are doing so only in moderation. American College Testing (ACT) English tests used in college admissions are still multiple-choice, although some smaller ACT testing programs use writing samples. The Scholastic Assessment Test (SAT) recently introduced an optional writing skill assessment that includes a *combination* of essay and multiple-choice questions. The Tests of General Educational Development (GED) Writing Skills Test (WST), the Advanced Placement English Language and Composition examination, and the Praxis tests for teacher certification all use *combinations* of essay and multiple-choice questions. Further details on these testing programs illustrate the approaches being used.

Despite these trends in some programs, however, it is also clear that other testing programs are not joining this trend or are doing so only in moderation.

Statewide Assessment Programs. According to the June 1995 annual report, *The Status of State Student Assessment Programs in the United States*, and its associated database, published by the Council of Chief State School Officers (CCSSO, 1995a, 1995b) and the North Central Regional Educational Laboratory (NCREL), 47 states have assessment programs. The number of states using performance-based assessments has grown from 17 in 1991-92, to 23 in 1992-93, to 25 in 1993-94. There is a clear trend toward writing samples, criterion-referenced testing, and alternative assessments, and away from norm-referenced multiple-choice assessments, in some states. Writing samples are used in 38 states. The number of states using writing portfolios remained constant at seven over this same period, however. Seventeen states use combinations of multiple-choice and performance tasks, while seven states use only multiple-choice assessments and two states use only alternative assessments coupled with writing samples. States

report that the major purposes of their assessments are improving instruction (43 states), school performance reporting (41 states), program evaluation (37 states), and student diagnosis (26 states). Only 17 states use their assessments for high school graduation, only 12 use them for school accreditation, and only two for teacher evaluation. Students are assessed most often in grades 4, 8, and 11. A special problem in statewide assessment is the requirement, often by law, that the same assessment be used for both accountability and instructional purposes.

A special problem in statewide assessment is the requirement, often by law, that the same assessment be used for both accountability and instructional purposes.

The National Assessment of Educational Progress (NAEP). The NAEP writing measure has always been a direct measure of writing, and it did not change much from the 1970s up until 1992. In 1992, a new framework was developed that increased the administration time from 15 minutes to 25 minutes and included 50-minute prompts at grades 4 and 8. Additionally, a planning page was included after each prompt, and the scoring rubrics were increased from 4 to 6 levels. Finally, a writing portfolio was introduced in 1992 to provide an in-depth look at classroom writing (NAEP, 1994a, 1994b). It is important to note that the NAEP testing program is intended to produce *aggregate* data for national or state samples of students, and thus does not encounter the same kinds of problems as testing programs aimed at individual assessment.

Medical College Admission Test (MCAT). The MCAT introduced a writing skill assessment consisting of two 30-minute essays in 1991. The essay topics present a brief quotation; the examinee is asked to explain the meaning of the quotation and then answer specific questions about it. Since the MCAT is an all-day battery of tests with about six hours of actual testing time, the new writing skill assessment represents only about one-sixth of total testing time.

Graduate Management Admission Test (GMAT). The GMAT writing assessment was introduced in 1994. Similar to the MCAT writing assessment, the GMAT Analytical Writing Assessment (AWA) consists of two 30-minute writing tasks. One of the 30-minute writing tasks is called "Analysis of an Issue" and the other "Analysis of an Argument." The AWA is designed as a direct measure of an examinee's ability to think critically and communicate ideas. The issues and arguments are often presented as quotations, as in the MCAT, but the quotations are longer. The responses are scored holistically, and copies of responses are included in admissions materials. The writing assessment represents

about one-fourth of total GMAT testing time. The GMAT also includes, as part of the GMAT Verbal Reasoning measure, a 25-minute sentence correction test in multiple-choice format.

Graduate Record Examination (GRE). The GRE plans to introduce at least one 45-minute essay in 1999. An additional 30-minute essay will be included if field tests now underway support it. The writing measure will represent somewhere between one-third and one-half of total testing time, depending on the outcomes of the field tests.

Law School Admission Test (LSAT). The LSAT introduced a 30-minute writing sample in 1982. Rather than being scored, however, the LSAT writing sample is reproduced and included with admissions materials for each law school applicant.

American College Testing (ACT) English Test. The ACT English test is a 45-minute multiple-choice test with 75 questions. The test assesses understanding of the conventions of grammar, sentence structure, and punctuation, as well as strategy, organization, and style. Five passages are presented, and each passage has several questions associated with it. Some questions refer to the passage as a whole, while other questions are about underlined words or phrases. Three scores are reported: a total score, a subscore on usage and mechanics, and a subscore on rhetorical skills. No free-response writing is required.

Scholastic Assessment Test (SAT). In 1994, the SAT was revised to include two parts: SAT I Reasoning and SAT II Achievement. The SAT II assessment is separate from SAT I and may or may not be required by institutions to which students are seeking admission. SAT II Writing, which is administered five times per year, consists of a 20-minute essay and a 40-minute multiple-choice test based on sentences and brief passages. A total score is reported, as are scores for both the essay and the multiple-choice tests.

Tests of General Educational Development (GED). The GED tests are used to grant a high school diploma to adults who did not complete high school. The Writing Skills Test (WST) of the GED consists of 50 multiple-choice questions and a single essay. Examinees have two hours to complete the WST, and they are advised to use 75 minutes for the multiple-choice questions and 45 minutes for the essay. The

essay is scored by two trained readers. The essay and multiple-choice sections are weighted (.36 and .64, respectively) and then scaled to form a single composite score.

Advanced Placement English Language and Composition (AP/EL&C). The AP/EL&C examination consists of 60 multiple-choice questions and three essays. Each of the three essays is read and scored by a different reader, and the total essay score is weighted 60 percent (versus 40 percent for the multiple-choice portion) in a composite grade reported on a 1-5 scale.

Praxis. The Praxis teacher certification tests, initiated in 1992 as a successor to the National Teacher Examinations (NTE), include a writing test with 45 multiple-choice questions and one 30-minute essay. The multiple-choice writing test can be taken in either a paper-and-pencil mode or as a computer-based test. If the computer-based multiple-choice test is taken, the essay may be written either with paper and pencil or with a word processor. The multiple-choice questions test understanding of subject-verb agreement, noun-pronoun agreement, correct verb tense, parallelism, clarity, and other conventions of standard written English. The prompts for the essay pose questions of relevance to teachers.

Most of the arguments for performance-based testing relate to hypothesized effects of testing on curriculum and instruction, but there are other arguments as well.

Arguments for Free-Response Writing Tasks

The arguments for the use of free-response writing tasks in the assessment of writing skill are essentially those of the performance testing movement, in which writing is often a focus. Most of the arguments for performance-based testing relate to hypothesized effects of testing on curriculum and instruction, but there are other arguments as well. The various arguments often overlap and, at times, seem to be the same argument using different terminologies.

Decomposition/decontextualization. One of the more elaborate arguments is that of Resnick & Resnick (1990). The argument begins by stating that two key assumptions, *decomposability* and *decontextualization,* underlie traditional standardized testing. The assumption of decomposability is that thought can be fractionated into independent pieces of knowledge, as in multiple-choice tests of writing skill when brief, independent problems in sentences are posed rather than a requirement for actual composition. The decontextualization assumption is that competence can be assessed "in a context very different from that in which it is practiced and

used," [p. 71] as in standardized testing in writing for which examinees are asked to edit the writing of someone else. Essay assessments in writing for which judges evaluate performances are given as an example of performance assessments, which are seen as a means for releasing educators from "the pressure toward fractionated, low-level forms of learning rewarded by most current tests . . ." [p. 78].

Systemically valid tests "induce curricular and instructional changes in education systems" and "foster the development of the cognitive traits that the tests are designed to measure"

Systemic Validity. Systemically valid tests "induce curricular and instructional changes in education systems" and "foster the development of the cognitive traits that the tests are designed to measure" (Frederiksen & Collins, 1989). Such tests are described as being direct (as opposed to indirect) and require subjective judgment in the assignment of scores. As in the decomposition/decontextualization argument, tests that emphasize isolated skill components, rather than higher-level processes, are seen as having a negative impact on instruction and learning. Free-response essay tests of writing skill are cited as examples of systemically valid tests.

Authenticity. Wiggins (1989, 1993) has argued that traditional standardized testing is not "authentic" (e.g., Wiggins, 1989, 1993). This another way of saying that tests are often not representative of real-life tasks. A multiple-choice test of verbal analogies, for example, is easily shown to involve tasks that are not encountered in everyday life in either school or work. However, Wiggins (1993) also observes that very brief essay tests are not authentic because in real life one has more time to write:

> *"Thus whatever assessors are testing in a 20-minute essay, it is certainly not the ability to write. As those of us who write for a living know, writing is revision, a constant returning to the basic questions of audience and purpose . . ." [p. 208].*

Authentic tasks are seen by Wiggins as those that are non-routine and multistage, that require the student to produce a high-quality product, and that are transparent in the sense that the student knows what to expect and can prepare for them.

Authentic tasks are seen by Wiggins as those that are nonroutine and multistage, that require the student to produce a high-quality product, and that are transparent in the sense that the student knows what to expect and can prepare for them. Dwyer (1993) observed considerable confusion about just what "authentic assessment" is perceived to be, however.

Teaching to the Test. Archbald & Porter (1990) described two main lines of argument advanced in criticisms of educational testing. The first is that testing adversely affects curriculum and instruction:

> *"Mandated student testing is conducted almost exclusively using facts and skills-dominated multiple-choice tests. Because there is accountability pressure for schools to achieve high test scores . . . teachers are forced to 'teach to the tests' — that is to shape their curriculum and instruction around the goal of developing students' test-taking abilities" [p. 34].*

This argument is continued by noting that what the tests do not measure does not get taught. Creativity, depth of understanding, integration of knowledge, ill-structured problem solving, and communication, for example, are not often included in tests and thus do not get taught. Another way of referring to this argument is to say that "the assessment tail wags the curriculum dog" (Swanson, Norman, and Linn, 1995).

Teacher Professionalism. A second argument against traditional testing given by Archbald and Porter (1990) is that it erodes teacher professionalism. When tests are used to make judgments about teacher or school quality, as well as promotion or retention of students, they exert a strong influence on what is taught and undermine teachers' pedagogical autonomy and feelings of professional worth. Similarly, White (1994), in the context of writing skill assessment, argues that holistic scoring of writing samples (as contrasted to multiple-choice tests) requires an "interpretive community" of teachers of writing "whose work is made meaningful by a joint social purpose" [p. 281].

A sophisticated line of argument supporting performance testing comes from the growing fields of cognitive and instructional psychology and from the testing establishment itself . . .

Cognitive Science. A sophisticated line of argument supporting performance testing comes from the growing fields of cognitive and instructional psychology and from the testing establishment itself, which in recent years has been more influenced by these academic disciplines. Part of this support comes from increasing interest in diagnosis and feedback. Nichols (1994) describes a new type of assessment termed "cognitively diagnostic assessment" (CDA) that requires analysis of "processes and knowledge structures involved in performing everyday tasks." Although CDA is responsive to some of the same educational concerns as performance-based testing, Nichols does not support all performance-based testing:

". . .scores on new performance-based or authentic assessments often provide little more information than traditional assessments to guide specific instructional decisions. Performance-based or authentic assessments may well consist of tasks that are more representative of some intended domain; however, these assessments continue to be developed and evaluated with an eye toward the same criterion — estimating a person's location on an underlying latent continuum. In either case, scores indicate no more than the need for additional instruction" [p. 578].

In writing assessment, for example, free-response essay tests, when scored holistically, do not provide sufficient diagnostic feedback to inform instruction.

In writing assessment, for example, free-response essay tests, when scored holistically, do not provide sufficient diagnostic feedback to inform instruction. CDA models focus on patterns of responses rather than average or total scores. The focus on patterns of responses is also reflected in arguments advanced by Mislevy (1993) for a new paradigm for assessment and by those interested in the psychology of problem solving (e.g., Snow & Lohman, 1989).

Cautions from the Measurement Community

As performance assessments have begun to be widely implemented, a number of articles have appeared about the standards of quality that performance tests should satisfy and about validity, reliability, comparability, fairness, and other measurement issues.

As performance assessments have begun to be widely implemented in statewide assessments and in national admissions testing, a number of articles have appeared in educational measurement journals posing questions about the standards of quality that performance tests should satisfy and about validity, reliability, comparability, fairness, and other measurement issues (e.g., Dunbar, Koretz, & Hoover, 1991; Linn, Baker, & Dunbar, 1991; Linn, 1994; Linn & Burton, 1994; Mehrens, 1992; Messick, 1994a, Messick, 1994b; Messick, 1995; Brennan & Johnson, 1995; Green, 1995; Bond, 1995). The following sections discuss these and other measurement concerns with a focus on writing skill assessment. Note that assessments of content knowledge, through the use of writing, are excluded from this discussion.

Content. What should be the content of a writing skill assessment? Ideally, the content of such an assessment might be based on models of writing skill developed from protocol analyses in which subjects are observed and asked to think aloud about what they are doing as they respond to a writing assignment. One of the most extensive writing model developments has been conducted by Hayes &

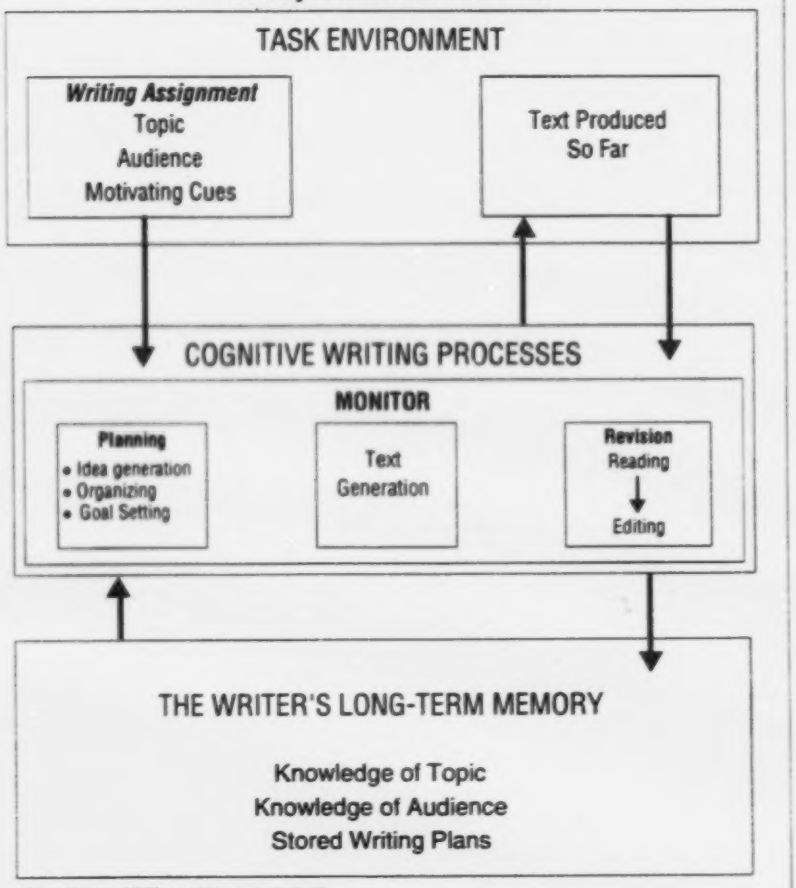
Flower (1980), Flower & Hayes (1981), and Hayes (1996) after many years of protocol analysis. The latest version of this model is shown in Figure 1. The central components of this model are the "cognitive writing processes," viz., planning, text generation, and revision. Other models of writing skill are in general agreement with the Hayes and Flower model. For example, Collins and Gentner (1980) separate writing into idea production and text production. Idea production occurs by keeping a journal of interesting ideas, obtaining ideas from others, reading books and other source materials, and attempting to explain one's ideas to someone else. Text production includes initial drafting, revision, and editing.

The central components of this model are the "cognitive writing processes," viz., planning, text generation, and revision.

Because writing assessments are usually constrained by a number of factors including the time available, security considerations, cost and reporting requirements, they rarely completely cover the domain of skills indicated by the writing

Because writing assessments are usually constrained by a number of factors . . . they rarely completely cover the domain of skills indicated by the writing models.

Figure 1
The Hayes and Flower Model



(From Hayes, 1996, used by permission)

models. Giving examinees advance notice of the topic to be written on (so that they will have time to generate ideas) cannot usually be allowed because that would create a test security problem. Quite often, a choice is made between drafting and revision because of time constraints. NAEP, MCAT, GMAT, LSAT, and some statewide assessments have opted for drafting only, with essays, letters, or some other free response as the vehicle. The planned GRE writing assessment also intends to assess drafting skills. Most often the free-response tasks assigned must be produced in a relatively brief period of time, although NAEP and several statewide assessments use writing portfolios in addition to brief assignments.

When a balance of assessment types is used, the amount of time available for each type becomes even more limited. In the SAT II writing assessment, for example, only 20 minutes of the one-hour test is allowed for the essay. If the entire test were free-response only, there would be no time for revision and editing tasks. That is the choice that has been made for the NAEP, MCAT, and GMAT writing tests, and planned for the GRE writing measure. This dilemma is partly resolved by recognizing that, "no matter how realistic a performance-based assessment is, it is still a simulation, and examinees do not behave in the same way they would in real life" (Swanson, Norman, and Linn, 1995). Nevertheless, the GMAT writing measure introduced in 1994 is already being criticized in the management literature for requiring only drafting, among other criticisms (Rogers & Rymer, 1995a, 1995b).

A more important source of error in free-response writing skill assessments is that due to the sampling of tasks for the writing assignments.

Reliability. The reliability of free-response writing examinations is often reported in terms of inter-rater reliability. This can be simply the correlation of scores assigned by two different raters to the same set of free responses. Unfortunately, such correlations always inflate the estimate of reliability, because only the error introduced by the raters is included. A more important source of error in free-response writing skill assessments is that due to the sampling of tasks for the writing assignments. What this means is that error is introduced because the examinee may be allowed to write on only a single topic chosen by the examiner. This topic may be an easy one for some examinees, because they may have recently written or thought about it, but a difficult topic for other examinees because they may never have thought about it. Accordingly, while the inter-rater correlation may be as high as .90, the score reliability for the same assessment may be only .50 because of the error introduced by topic sampling. See Reckase

(1995b) for further explanation of this problem. Dunbar, Koretz, and Hoover (1991) reviewed reliability studies for common types of assessments and showed that essay tests of the type most often used (one topic, two readers) have score reliabilities of around .50, on average. It is often more useful to examine the standard error of measurement, derived from reliability, as recommended by Linn (1994) and Linn & Burton (1994). The standard error of measurement is especially important when using assessments to make classification decisions such as pass or fail.

. . . essay tests of the type most often used (one topic, two readers) have score reliabilities of around .50, on average.

Better reliabilities can be obtained by using more tasks and more raters. The MCAT and GMAT writing assessments, with two tasks of about 30 minutes each, and each task scored independently by two different raters, will produce much better reliabilities than the usual free-response assessment. Breland et al. (1987) estimated that essay assessments of this type could yield score reliabilities in excess of .70. Nevertheless, Linn (1994) suggests that reliabilities even as high as .80 can be problematical.

The only way that the reliability problem of free-response tasks can be resolved is through the use of multiple samples written by the same examinee, with each scored independently by multiple raters.

The only way that the reliability problem of free-response tasks can be resolved is through the use of multiple samples written by the same examinee, with each scored independently by multiple raters. Reckase (1995a) shows that to approximate a .80 reliability, a writing assessment of five different samples is required, and the components of the assessment need to be similar rather than disparate. Breland et al. (1987) estimated that, for a hypothetical portfolio of six essays, each scored by three different raters, a score reliability of .88 could be attained. These high reliabilities are obtained only when all examinees write on the same topics under the same conditions and when the scoring is conducted by the same raters. That is, the tasks and administrative conditions (timing, scoring) are standardized. Writing portfolios, for which examinees submit writing samples of their own choice written on widely different topics and under varying conditions, are not likely to attain such high levels of reliability. Some statewide assessments using writing portfolios, notably one initiated in Vermont in 1988, have encountered serious problems with reliability (Koretz, Stecher, Klein, & McCaffrey, 1994). Better reliabilities were obtained with the NAEP writing portfolio, however (Gentile, 1992).

The reliability of writing assessments can be increased by combining assessment types (essay and multiple-choice) as is done for the SAT II, GED, Advanced Placement, and Praxis writing assessments.

The reliability of writing assessments can be increased by combining assessment types (essay and multiple-choice) as is done for the SAT II, GED, Advanced Placement, and Praxis

writing assessments. The GED writing assessment, with a single 45-minute essay and 50 multiple-choice questions, yields a reliability of about .87 (Patience & Swartz, 1987; Lukhele & Sereci, 1995; Wiley & Sireci, 1994). The GED essay is scored by two trained readers on a six-point scale. The Advanced Placement English Language and Composition examination, with three free-response tasks each scored by a single reader, and a 100-item multiple-choice test, produces reliabilities in the .78 to .90 range (College Board, 1988).

As a final note on reliability of writing assessments, it is important to point out that the NAEP assessments, as aggregations of data intended for the assessment of national trends, do not have the same reliability problems as do assessments intended to produce scores for individual examinees. Similarly, statewide assessments intended for decision making at the school or district level can resolve reliability problems by pooling data for an entire school or district as well as across years, as is done for the Kentucky Instructional Results Information System (KIRIS) analyzed by Haertel (1994).

Predictive Effectiveness. The predictive effectiveness of writing skill assessments, important for admissions tests, is related to reliability. High reliability is a necessary but not a sufficient condition for predictive effectiveness. But even a highly reliable test will not predict well if it is assessing the wrong skills. To assess predictive effectiveness, test scores are correlated with some later outcome, such as grades in English courses or freshman grade point average (GPA). Bridgeman (1991), for example, analyzed the effectiveness of multiple-choice and essay assessments of writing skill for predicting college freshman GPA and obtained median correlations across 21 colleges of .30 for the multiple-choice assessment and .16 for the essay. The essay assessment did not add incrementally to the predictive effectiveness possible using multiple-choice tests alone. When English composition course grades, rather than GPA, have been used as the criterion, however, incremental predictive validity for essays has been observed (Breland and Gaynor, 1979).

To assess predictive effectiveness, test scores are correlated with some later outcome, such as grades in English courses or freshman grade point average (GPA).

Much higher predictive correlations can be obtained if, instead of grades, scores on performance tests of writing are predicted. Breland et al. (1987) obtained high correlations for predicting writing performance. A writing performance assessment consisting of five essays, each scored by three

different readers, correlated well with a number of predictors. A 30-minute multiple-choice test of grammar and sentence structure questions correlated .70 with the writing performance assessment. When the same 30-minute multiple-choice test was combined with a single 45-minute essay assessment with two independent readings, the predictive correlation increased to .77. Two essays alone, when used to predict scores on a writing performance assessment consisting of four essays, yielded predictive correlations in a range from .61 to .75. It seems clear from these analyses that good predictions of writing performance can be made using multiple-choice tests alone, essays alone, or combinations of essays and multiple-choice tests. When essays are used alone, however, it is preferable to use more than a single essay and more than a single reader of each essay.

It seems clear from these analyses that good predictions of writing performance can be made using multiple-choice tests alone, essays alone, or combinations of essays and multiple-choice tests.

Fairness. In writing skill assessment, fairness issues have tended to focus on differences in gender, race, and language. There is much evidence to suggest that women tend to write better than men, on average, and that this advantage is more pronounced in free-response assessments than it is in multiple-chcice assessments. Members of racial minority groups, as well as linguistic minorities, tend to score lower than non-minorities on all types of writing skill assessments (see, e.g., Breland & Griswold, 1982; Klein, 1989; Murphy, 1982; Petersen & Livingston, 1982). A special problem encountered by linguistic minorities occurs when writing skill assessments consume a large proportion of the testing time of a more comprehensive assessment.

Contrary to popular opinion, however, performance assessments do not necessarily result in better outcomes for minorities than multiple-choice assessments. Bond (1995) cites a number of papers suggesting that, for NAEP, extended-response essays resulted in mean differences between African Americans and Whites that were equal to those for the multiple-choice reading assessment. After correcting for unreliability, the mean differences actually exceeded those found on the multiple-choice reading assessment. Klein (1989) showed that increased essay testing on the California bar exam did not reduce the differences in passing rates between White and minority groups.

Contrary to popular opinion, however, performance assessments do not necessarily result in better outcomes for minorities than multiple-choice assessments.

Women have been shown to perform better on essay examinations than would be expected from their scores on multiple-choice tests of writing skill (e.g., Breland and Griswold, 1982), and in the Klein (1989) bar exam study, the passing rate for women increased when the amount of essay testing was increased.

Consequences. Messick (1995) observes that it is important to collect evidence of both positive and negative consequences of performance assessments. If the promised benefits to teaching and learning occur, then this is evidence in support of the validity of performance assessments. If such benefits do not occur, or if there are negative consequences, that is also important to document. If negative consequences result, it is important to determine their causes. If some examinees receive low scores because something is missing from the assessment, this is evidence of construct under-representation. That is, for example, if a writing assessment consists only of questions about how to revise text, and does not allow an examinee to demonstrate an ability to produce text, then the construct as defined by the Hayes and Flower model of writing is underrepresented. Additionally, low scores should not occur because the assessment contains irrelevant questions. In writing assessment, for example, an essay prompt on a topic that examinees are unlikely to be familiar with could affect performance unfairly.

Unfortunately, evidence on the consequences of new forms of assessment is rarely assembled.

Unfortunately, evidence on the consequences of new forms of assessment is rarely assembled. In the health professions, for example, Swanson, Norman, and Linn (1995) could find only two examples of systematic research on the impact of changes in examinations. One reason for the failure to conduct research on consequences is that it is difficult (Linn, 1994). It may require a number of years for a new assessment to produce observable changes in the behaviors of students or teachers, or the changes may be so gradual that they are not easily detected.

In writing skill assessment, comparability is a particularly troublesome problem because individual free-response tasks are quite often not comparable.

Comparability. In order to make comparisons of assessments from year to year or from administration to administration, the assessments must mean the same thing on different occasions. This means that they must be of comparable content and of comparable difficulty. In writing skill assessment, comparability is a particularly troublesome problem because individual free-response tasks are quite often not comparable. Comparability problems are alleviated to some extent through the use of multiple tasks, as in NAEP and some statewide assessments, or through the combination of free-response tasks with multiple-choice items, as is done for SAT II, Advanced Placement, the GED, and Praxis.

Comparability problems are alleviated to some extent through the use of multiple tasks, as in NAEP and some statewide assessments, or through the combination of free-response tasks with multiple-choice items, as is done for SAT II, Advanced Placement, the GED, and Praxis.

To ensure comparable content requires careful attention to test specifications (Green, 1995). With traditional

multiple-choice tests, test specifications are made comparable across testing occasions by balancing the number of items of each type. With a large number of items, balancing test specifications is not difficult. In writing skill assessment using free responses, however, the number of tasks is usually quite small, and each task may require from 20 minutes to an hour of time. As a result, comparability of content may be difficult to maintain. It is of course essential, in addition, to control the exposure of free-response tasks so that the content does not become known prior to a test administration (Mehrens, 1992). The scoring of free-response tasks must be carefully controlled across administrations by use of the same scoring rubrics and reader training from year to year or from administration to administration.

Some free-response writing tasks do not appeal to some examinees, and the resulting examinee-task interaction tends to lower reliability.

Finally, comparability is also affected by the reliability of a test, since the less reliable the test the less reliable will be equating across forms and occasions (Green, 1995). Some free-response writing tasks do not appeal to some examinees, and the resulting examinee-task interaction tends to lower reliability. A number of studies have demonstrated that examinee-task interactions are a major source of error in essay examinations (Breland et al., 1987; Brennan & Johnson, 1995; Coffman, 1966).

It is important to note that, despite the considerable difficulties resulting from comparability problems in individual performance assessment, aggregate assessments such as NAEP, statewide, districtwide, and schoolwide assessments can often be made comparable through careful sampling designs and the rotation of free-response prompts (Green, 1995; Linn, Baker, & Dunbar, 1991).

Cognitive complexity. Proponents of performance testing often note the need for "higher-level" assessments or for "ill-structured problems" (e.g., Frederiksen, 1984; Resnick & Resnick, 1990). Unfortunately, not all examinees can solve such difficult problems and, as a result, cannot be accurately evaluated by them. An example in writing skill assessment is when a prompt is a quotation of some type. The quotation may be a famous one such as Descartes' "I think, therefore I am." The examinee's task is to write a well-organized and unified essay in response to a question about such a quotation. In the MCAT, for example, the first task is to explain what the quotation means. If the examinee does not know what the quotation means, which seems quite likely, then it is difficult to write anything at all. That is, the requirement of cognitive complexity interacts with the consequences

discussed earlier. Messick (1994b) observed that "low scores should not occur because the measurement contains something irrelevant that interferes with the affected persons' demonstration of competence." In standardized testing, it has long been recognized that a range of difficulties in questions is needed to obtain good assessments for all examinees.

In standardized testing, it has long been recognized that a range of difficulties in questions is needed to obtain good assessments for all examinees.

Realism. It is often assumed that an essay examination is unquestionably a realistic representation of a real-life task. A brief 20- to 45-minute essay, in response to a topic previously unknown to the examinee, is hardly realistic. Truly realistic assessments of writing skill would require several samples of writing produced without severe time constraints and evaluated by multiple judges. In most writing assessments the time available is limited, and administrative and scoring costs must be controlled. Most writing skill tests, therefore, can only be simulations of the skill being assessed. From this perspective, a brief essay, even if on an impromptu topic, is a useful simulation of writing. Likewise, a brief editing task is also a useful simulation of a real-life task even if the writing to be edited was written by someone other than the examinee.

Cost and Efficiency. Writing portfolios and multiple-choice tests of writing represent two extremes on a cost/efficiency continuum. There can be little doubt that a carefully designed writing portfolio is a reliable and valid assessment of writing skill. Moreover, such a portfolio should have a positive impact on the curriculum. Nevertheless, a writing portfolio would not be the optimum assessment for all purposes because of its cost and because of the amount of time required to provide feedback to the examinee or to the educational system. As used in NAEP for a representative national sampling of students, a writing portfolio seems appropriate, as it can be for some statewide assessments. For individual assessment, as in college and graduate school admissions, a writing portfolio would be excessively expensive. The time required to develop and report scores would not be compatible with timing requirements in the admissions cycle of events.

Truly realistic assessments of writing skill would require several samples of writing produced without severe time constraints and evaluated by multiple judges.

At the other extreme of this cost/efficiency continuum, the multiple-choice test of the conventions of standard written English seems appropriate for some purposes but not for others. If the test has been demonstrated to be both reliable and valid for the purpose intended (e.g., the prediction of writing performance in college or graduate school),

then it would seem to be a likely candidate for use in those situations. Of course, all considerations described above, including fairness and consequences, need to be evaluated as well. It may be that the addition of a brief essay to the multiple-choice test will enhance its validity or fairness and have positive impact on the educational system.

Between the extremes of the writing portfolio and the multiple-choice test are other options, including the use of multiple brief writing tasks as in the MCAT and the GMAT.

Between the extremes of the writing portfolio and the multiple-choice test are other options, including the use of multiple brief writing tasks as in the MCAT and the GMAT. Here, the disadvantages of relatively low reliability and incomplete content coverage (with no revision or editing) may be offset by considerations of fairness and consequences while cost and efficiency, while not optimum, are acceptable.

Conclusion

The performance testing movement has had a positive impact on writing skill assessment practices.

The performance testing movement has had a positive impact on writing skill assessment practices. Essays and other free-response tasks are now being used much more widely, and their use helps to make writing assessments more representative of real-world writing by helping to cover more of the domain involved in actual writing. But it must be remembered that the writing tasks used in assessments, for the most part, can only be *simulations* of real-world writing. For that reason, free-response writing tasks may be no more authentic than any other kind of writing assessment. Some writing portfolios may closely approximate real-life tasks, but such portfolios are rare, and they are not suitable for all assessment purposes.

Another thing to remember is that there is also a positive side to the decomposition and decontextualization of writing skills. Most decomposed and decontextualized skills are much easier to teach and learn than writing itself. And these skills are important in their own right, even if they do not make a person a good writer. Take, for example, simple writing problems such as vague pronoun references, parallel structure in sentences, and transitions from one sentence to another. It is relatively easy to teach and learn about these simple kinds of writing problems, and they can be quite important in a person's life (say, in writing a letter of application for a job). Recognizing such writing problems in the writing of others is no different than recognizing them in one's own writing. Many other examples could be cited of decomposed and decontextualized writing skills that are important.

. . . evidence has yet to be assembled to show that free-response writing skill assessment will improve the writing skill of the nation or that other kinds of writing skill assessment have diminished these skills.

Finally, we need to remember that evidence has yet to be assembled to show that free-response writing skill assessment will improve the writing skill of the nation or that other kinds of writing skill assessment have diminished these skills. Perhaps the 1998 NAEP writing assessment, when compared to the 1992 NAEP writing assessment, will show an improvement. Until that occurs, or until other evidence shows that the nation's writing has improved, it seems best to be cautious about radical changes in our writing skill assessments.

The types of writing skill assessment that are most appropriate depend upon the purpose of the assessment.

The types of writing skill assessment that are most appropriate depend upon the purpose of the assessment. If the purpose is to gauge trends in national abilities, as in NAEP, or trends in states and districts, free-response writing assessments clearly seem to be appropriate. Reliability is much less of a problem because of the aggregation of data at the national, state, or district level. For purposes of high-stakes individual assessment, as for admissions decisions for college or graduate school, reliability problems are formidable, however, and caution needs to be exercised. One approach to handling the reliability problem of free-response tasks is to combine both free-response and multiple-choice tasks to make up the assessment, as is done for the SAT II writing assessment, Advanced Placement, Praxis, and the GED assessment. Another approach is to use multiple free-response tasks, as is done for MCAT and GMAT examinations, and which is being considered for the GRE writing assessment. Cost and other practicalities will usually limit the number of writing tasks used to two or three. The effects on reliability and validity of such limitations need to be examined.

Long experience with performance-based assessments in the health professions supports a blend of assessment methods.

Long experience with performance-based assessments in the health professions supports a blend of assessment methods (Swanson, Norman, and Linn (1995). A similar conclusion was reached by Ackerman & Smith (1988) through a factor analysis of both essay tests and multiple-choice tests of revision skills. Miller & Crocker (1990) in their review of validation methods for writing assessment, came to the same conclusion:

> *"Thus, it seems that when interested in providing a complete description of writing ability, both direct and indirect writing assessment are needed" [p. 292].*

From a domain coverage perspective, the use of combined methods seems more likely to cover both drafting and revision skills.

What are the prospects for the future? Technology seems to be an important key to the future of writing skill assessment. Already, for Praxis and other assessments, writing samples are collected by computer and thus are available for analysis, transmission, and evaluation using computer-based technologies. The inefficiencies of collecting writing samples on paper and having them evaluated by experts should be much less of a problem in the future. Samples can be transmitted to experts electronically for evaluation in their homes or offices without the need for travel to a central facility for scoring. Thus the future of writing skill assessment would appear to be one of increasing acceptability of performance tasks, even portfolios, because of the efficiencies that will be available through technology. Multiple-choice testing will still be useful, however, though "bubbles" on answer sheets will probably be replaced by mouse clicks on the computer, as they have been on the College Board's Computerized Placement Tests. It is quite likely that many editing tasks will be conducted by *constructing responses* rather than clicking on an answer. The computer should also make it possible to provide more diagnostic feedback to students and teachers than is currently possible.

Technology seems to be an important key to the future of writing skill assessment.

References

Ackerman, T. A., & Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. *Applied Psychological Measurement, 12* (2), 117-128.

Archbald, D. A., & Porter, A. C. (1990). A retrospective and an analysis of roles of mandated testing in education reform. Report prepared for the Office of Technology Assessment. PB92-127596. Springfield, VA: National Technical Information Service.

Bond, L. (1995). Unintended consequences of performance assessment: Issues of bias and fairness. *Educational Measurement: Issues and Practice, 14* (4), 21-24.

Breland, H. M., Camp, R., Jones, R. J., Rock, D., & Morris, M. (1987). *Assessing writing skill*. New York: College Entrance Examination Board.

Breland, H. M., & Gaynor, J. (1979). A comparison of direct and indirect assessments of writing skill. *Journal of Educational Measurement, 16* (2), 119-128.

Breland, H. M., & Griswold, P. A. (1982). Use of a performance test as a criterion in a differential validity study. *Journal of Educational Measurement, 74*, (5), 713-721.

Brennan, R. L., & Johnson, E. G. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice, 14* (4), 9-12, 27.

Bridgeman, B. (1991). Essays and multiple-choice tests as predictors of college freshman GPA. *Research in Higher Education, 32*, (3), 319-331.

Coffman, W. E. (1966). On the validity of essay tests of achievement. *Journal of Educational Measurement, 3* (2), 151-156.

College Board. (1988). *Technical Manual for the Advanced Placement Program*. New York: College Entrance Examination Board.

Collins, A., & Gentner, D. (1980). A framework for a cognitive theory of writing. In L. W. Gregg and E. R. Steinberg (Eds.), *Cognitive processes in writing*. Hillsdale, NJ: Erlbaum.

Council of Chief State School Officers (1995a). *The status of state student assessment programs in the United States*. Annual Report. Oak Brook, IL: North Central Regional Educational Laboratory, June.

Council of Chief State School Officers (1995b). *State student assessment programs database*. Oak Brook, IL: North Central Regional Educational Laboratory, June.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Psychological Measurement, 4* (4), 289-303.

Dwyer, C. A. (1993). Innovation and reform: Examples from teacher assessment. In R. E. Bennett & W. C. Ward Jr. (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 265-289). Hillsdale, NJ: Erlbaum.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*, 365-387.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*, 193-202.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18* (9), 27-32.

Green, B. F. (1995). Comparability of scores from performance assessments. *Educational Measurement: Issues and Practice, 14* (4), 13-15, 24.

Haertel, E. H. (1994). Theoretical and practical implications. In T. R. Guskey (Ed.), *High stakes performance assessment: Perspectives on Kentucky's Educational Reform*. Thousand Oaks, CA: Corwin Press, Inc.

Hayes, J. R. (1996). A new model of cognition and affect in writing. In C. Michael Levy and Sarah Ransdell (Eds.), *The Science of Writing*. Mahwah, NJ: Erlbaum.

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. W. Gregg and E. R. Steinberg (Eds.), *Cognitive processes in writing*. Hillsdale, NJ: Erlbaum.

Klein, S. P. (1989). *Does performance testing on the bar examination reduce differences in scores among sex and racial groups?* The RAND Corporation. Unpublished manuscript.

Koretz, D., Stecher, Brian, Klein, S., & McCaffrey (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice*, *13*, (3), 5-16.

Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher*, *23*, (9), 4-14.

Linn, R. L., Baker, E. V., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *20* (8), 15-21.

Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice*, *13* (1), 5-8, 15.

Lukhele, R., & Sireci, S. G. (1995). Using IRT to combine multiple-choice and free-response sections of a test onto a common scale using a priori weights. Paper presented at the annual meeting of the National Council on Measurement in Education, April, San Francisco, CA.

Mehrens, W. A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice*, *11*, (1), 3-9, 20.

Messick, S. (1994a). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23* (2), 13-23.

Messick, S. (1994b). *Alternative modes of assessment, uniform standards of validity*. Research Report RR-94-60. Princeton, NJ: Educational Testing Service.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, *14* (4), 5-8.

Miller, M. D., & Crocker, L. (1990). Validation methods for direct writing assessment. *Applied Measurement in Education*. *3* (3), 285-296.

Mislevy, R. J. (1993). A framework for studying differences between multiple-choice and free-response items. In R. E. Bennett & W. C. Ward Jr. (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 75-106). Hillsdale, NJ: Erlbaum.

Murphy, R. J .L. (1982). Sex differences in objective test performance. *British Journal of Educational Psychology*, *52*, 213-219.

National Assessment of Educational Progress. (1994a). *NAEP 1992 writing report card*. Princeton, NJ: Educational Testing Service.

National Assessment of Educational Progress. (1994b). *NAEP 1992 trends in academic progress*. Princeton, NJ: Educational Testing Service.

Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research*, *64* (4), 575-603.

Patience, W., & Swartz, R. (1987). Essay score reliability: Issues in and methods of reporting the GED Writing Skills Test scores. Paper presented at the annual meeting of the National Council on Measurement in Education, April, Washington, DC.

Petersen, N., & Livingston, S.L. (1982). *English composition test with essay*. A descriptive study of the relationship between essay and objective scores by ethnic group and sex. (ETS Statistical Report No. SR-82-96). Princeton, NJ: Educational Testing Service.

Reckase, M. (1995a). Portfolio assessment: A theoretical estimate of score reliability. *Educational Measurement: Issues and Practice*, *14* (1), 12-14, 31.

Reckase, M. (1995b). The reliability of ratings versus the reliability of scores. *Educational Measurement: Issues and Practice*, *14* (4), 31.

Resnick, L. B., & Resnick, D. P. (1990). Tests as standards of achievement in schools. In *The Uses of Standardized Tests in American Education: Proceedings of the 1989 ETS Invitational Conference* (pp. 63-80). Princeton, NJ: Educational Testing Service.

Rogers, P. S., & Rymer, J. (1995a). What is the relevance of the GMAT Analytical Writing Assessment for management education? A Critical Analysis, Part 1. *Management Communication Quarterly, 8*, (3), 347-367.

Rogers, P. S., & Rymer, J. (1995b). What is the functional value of the GMAT Analytical Writing Assessment for management education? A Critical Analysis, Part 2. *Management Communication Quarterly, 8*, (4), 477-494.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational assessment. In R. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 263-331). Washington, DC: American Council on Education.

Swanson, D. B., Norman, R. R., & Linn, R. L. (1995). Performance-based assessment: Lessons from the health professions. *Educational Researcher, 24* (5), 5-11, 35.

White, E. M. (1994). *Teaching and assessing writing*. San Francisco: Jossey-Bass.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*, 703-713.

Wiggins, G. (1993). Assessment: Authenticity, context, and validity. *Phi Delta Kappan, 75*, 200-214.

Wiley, A., & Sireci, S. G. (1994). Determining the reliability of the GED Writing Skills Test using classical test theory. Paper presented at the annual meeting of the Northeastern Educational Research Association, October, Ellenville, NY.